単語連鎖書き換え規則を用いた連語書き換え機構

吉田雄太 宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

日本語形態素解析 Maja[1] では、翻訳処理、音声処理など様々な用途での利用を見越して、日本語文を辞書に収録された基本語（短単位）の組合わせに分割する。しかし、例えば機械翻訳では、格助詞相当の連語、助動詞相当の複合辞などのように長単位で扱ったほうが処理がしやすい。そのため、形態素解析の後処理として、単語連鎖列から連語となる部分を抽出し、長単位の単語に置き換える手法が必要とされる。

本稿では、形態素解析の後処理として、単語連鎖列から連語句となる部分を抽出し、その前後部や離散的な条件を考慮した上で、単語連鎖書き換え規則を用いて、長単位の単語に置き換える手法を提案し、その有効性を示す。

2 連語書き換え機構の必要性

「努力によって成功。」という文を構文解析する場合、「努力/に/よ/っ/て/成功/。」というようにまず形態素解析によって文を最小構成要素である単語に分割し、同時に各単語の品詞を決定する必要がある。しかし、このような単語列をそのまま構文解析の入力として解析をすると、「によって」が、'格助詞「に」+本動詞「よっ」+助動詞「て」'として扱われてしまい、「努力」という「場所」に「寄って」、「成功」という誤った構造を導き出してしまう。

Collocation Rewriting Mechanism
using Word String Rewriting Rules
Yuta Yoshida, Masahiro Miyazaki
Niigata University

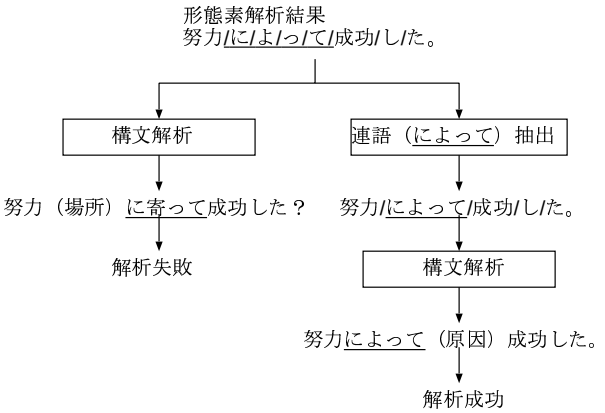

図１:連語解析機構の流れ

そこで、図１に示すように、形態素解析の結果から連語の部分を判別、抽出し「努力/によって/成功/。」という単語列に書き換える。その上で構文解析を行うことによって、「努力」という「原因」で、「成功」という正しい構造を導き出すことが可能になる。そのために、連語となりうる語の並びを持つ形態素解析結果に対し、その語の並びを連語として処理してよいかを判断し、連語として処理すべきものは形態素解析結果を書き換えて、構文解析に送る連語書き換え機構が必要とされる。

3 連語書き換え規則の形式

連語書き換え機構に用いる連語書き換え規則の形式を以下に示す。

連語書き換え規則の形式

$W_{\ i}(\mathrm{pos}), *, —, *, W_{\ 1}(\mathrm{pos}), *,$
［$W_1(\mathrm{pos}), —, W_j$］
$, *, W_{+1}(\mathrm{pos}), *, —, *, W_{+k}(\mathrm{pos})$
$\rightarrow W_{\ i}(\mathrm{pos}), *, —, *, W_{\ 1}(\mathrm{pos}), *,$
$W'(\mathrm{pos})$
$, *, W_{+1}(\mathrm{pos}), *, —, *, W_{+k}(\mathrm{pos})$
W:字面、pos:品詞

字面と品詞の組み合わせ $W(\mathrm{pos})$ が一つの単語に対応し、その単語の連鎖列によって [ ] の単語列を $W'(\mathrm{pos})$ に書き換えるようになっている。また、[ ] の前後の単語列は連語書き換え規則の適用の可否やその条件を表す。また、＊は長さ 0 以上の任意の単語列を表す。これによって離散的な単語連鎖条件にも対応できる。なお、これらの条件は省略も可能であり、その場合、無条件に連語書き換えを行う。また、品詞を示す pos には、否定条件などにも対応させているため、複雑な条件を簡潔に表記することが可能であり、ルールの追加も容易に行えるようになっている。

4 連語書き換えの適用例

例 1 :「によって (助詞相当)」

＊ (ダ型静詞、「に」に接続する副詞、場所を表す名詞を除く), ［に (格助詞), よ (五段動詞語幹),っ(五段動詞語尾), て (既定の助動詞)］, ＊（形式動詞を除く）
→によって (格助詞)

2で例として述べた、「によって」を扱う規則である。前方と後方の条件は「場所」に「寄って」という意味を持つ「によって」という解析結果を書き換えないために必要となる。以下の例では、A の文は「寄って」の意であり、B の文が「原因」を表す格助詞相当となる。

A. 右 によって いる。 (一般用法)
B. 努力 によって 成功。 (格助詞相当)

このルールの評価実験結果を表 1 に示す。

表 1 :「によって」の評価実験の結果

	A タイプの文	B タイプの文
総数	1	99
解析正解数	1	99

例 2 :「として (助詞相当)」

＊ (用言、助動詞、「と」に接続する副詞を除く), ［と (格助詞), し (サ変動詞), て (既定の助動詞)］, ＊（形式動詞を除く）
→として (格助詞)

役割を表す「として」を扱う規則である。条件は、「する」という意味で、「として」が扱われている場合のために用意されている。以下の例では、A の文は「する」の意であり、B の文が「役割」を表す格助詞相当となる。

A. 始まった として いる。 (一般用法)
B. 人 として 生きる。 (格助詞相当)

このルールの評価実験結果を表 2 に示す。

表 2 :「として」の評価実験の結果

	A タイプの文	B タイプの文
総数	15	85
解析正解数	11	85

5 おわりに

本稿では短単位に分割された形態素解析結果に対し連語書き換え機構を用いることにより、複雑な適用条件を持つ連語に対し効果的に書き換える機構を提案し、その有効性を示した。この枠組みを利用し、日本語的表現など翻訳が困難な語句を翻訳に適した語句に書き換える等価変換や、同形語判別、異なった形態素解析システムの出力結果への変換など様々な応用が期待される。

参考文献

[1] 尾島、宮崎:日本語形態素解析システムにおける部分的再試行機構の導入とその効果、情報処理学会第 58 回全国大会 1E-4(1999)